CHAPTER 18. REGRESSION AND CORRELATION.

Regression experiments. Consider a probabilistic experiment which has a value Y as its outcome, but where the experiment itself, and hence the probabilistic behavior of Y, is affected by the value assigned to some parameter X which is controlled by the experimenter. For each trial of the experiment, the parameter is given some chosen fixed value. This chosen value may vary from trial to trial. For example, the experiment might be randomly to choose an individual and then to give that individual a certain standard test, where Y is the score obtained on the test and X is the amount of time allowed to the individual for taking the test. Such a parameterized experiment is called a regression experiment. (The reason for the term "regression" is explained later in this chapter.)

Regression experiments occur widely and are frequent objects of statistical study. Typical examples include the following:

(a) A seed of a certain plant species is randomly selected and planted in a soil of standard composition. A special nutrient is added to the soil and the growth rate of the plant is observed. Here Y = observed growth rate, and X = concentration in the soil of the special nutrient.

(b) The experimenter makes n rolls of a single die and observes the sum of the numbers that appear. Here Y = observed sum, and $X = n$.

(c) The electrical conductivity of a certain metal is approximately measured at a precisely fixed temperature T.

Here Y = observed value of conductivity and $X = T$. (For a given T, observed values of Y may vary because of random experimental errors occurring in the procedure for measuring conductivity. Near the end of this chapter, we shall also consider situations where random experimental errors occur in measurements of X as well as in measurements of Y.)

For each chosen value of the parameter X in a regression experiment, the observed value Y may be viewed as a random variable. To simplify our exposition, we shall assume from now on that for each value of X, Y is a continuous random variable on $(-\infty,\infty)$. (This was not the case in (b) above.) We shall usually take X to be continuously variable on $(-\infty,\infty)$, but sometimes we shall have X with values on an interval such as $(0,\infty)$ or on some discrete set of real numbers such as the non-negative integers.

In a regression experiment, the parameter X is called the <u>independent variable</u>, and the corresponding random variable Y is called the <u>dependent variable</u>.

<u>Regression models</u>. Given a regression experiment, a <u>regression model</u> for that experiment is a rule which gives, for each value x of X, a probability density function $g_x(y)$ for the corresponding random variable Y. A regression model may hence be viewed as a family of density functions, one for each possible value of X. (As usual for density functions, we require, for each x, that $g_x(y) \geq 0$ and that $\int_{-\infty}^{\infty} g_x(y)dy = 1$.) Evidently, a regression model may be viewed as a function $g(x,y)$ of two variables, where $g(x,y) = g_x(y)$. This function is also

often written as $g(x,y) = g(y|x)$, and the latter notation is read as "the probability density for Y at y, given x."

We introduce the following notations. Let

$$N(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2} .$$

Thus $N(y;\mu,\sigma)$ is a normal density with mean μ and variance σ^2.

Let $C(y;m) = \dfrac{1}{\pi(1+(y-m)^2)}$.

Then $C(y;m)$ is a Cauchy density with median m.

Examples. We consider the following.

(1) $g(y|x) = C(y;3x)$. This is a regression model where, for each value x of X in $(-\infty,\infty)$, Y is a Cauchy variable whose median = $3x$.

(2) $g(y|x) = N(y;x^2,2)$. This is a regression model where, for each value x of X in $(-\infty,\infty)$, Y is a normal variable with mean x^2 and variance 4.

(3) $g(y|x) = N(y;5x-7,x)$. This is a regression model where, for each value x of X in $(-\infty,\infty)$, Y is a normal variable with mean $5x-7$ and variance x^2.

(4) $g(y|x) = N(y;5x-7,2)$. This is a regression model where, for each value x of X in $(-\infty,\infty)$, Y is a normal variable with mean $5x-7$ and variance 4.

Consider a regression model $g(y|x)$ where, for all values of X, Y has both expectation and variance. We can then define

$$E(Y|x) = \int_{-\infty}^{\infty} y\, g(y|x)\, dy .$$

For each fixed value x of X, $E(Y|x)$ is called the expectation of Y given x. Evidently $E(Y|x)$ is a function of x. The equation $y = E(Y|x)$ is called the regression of Y on X, and the curve in the XY plane which it defines is called the regression curve of Y on X (for the given model). In example (2), the regression of Y on X is $y = x^2$; in (3) and (4), it is $y = 5x - 7$. A model is said to be a linear regression model if its regression is linear. Thus (3) and (4) are linear regression models, but (2) is not. (Strictly speaking, (1) is not a linear regression model, since $E(Y|x)$ is not defined.)

Similarly, we can define the variance of Y given x as

$$V(Y|x) = \int_{-\infty}^{\infty} (y - E(Y|x))^2 g(y|x)\,dy \ .$$

If the density for Y in a given regression model is normal for every value of X, we say that the model itself is a normal regression model. Thus (2), (3), and (4) are all normal regression models.

Finally, for a given regression model, if $V(Y|x)$ remains constant as x varies, we say that the model is homoscedastic. Examples (2) and (4) are homoscedastic. Otherwise (as in (3)) we say that a model is heteroscedastic.

In what follows, we shall confine ourselves to regression models which are normal, linear, and homoscedastic. We shall refer to such models as simple normal linear regression models or, more briefly, as SNLR models.

Prevalence of SNLR models. If, in a given parametrized experiment, we have reason to expect that the mean of Y changes linearly with the assigned value for X, we may wish to assume a linear regression model. If, in addition, for each value for X, th variation in Y about its mean value appears to be the result of a number of small and largely independent influences, we may, by the Central Limit Theorem, wish to assume a normal model. Finally, if we believe that the physical sources of this variation in Y about its mean are largely independent of the specific value assigned to X, we may wish to assume a homoscedastic regression model. This combination of circumstances arises frequently in nature and in statistical practice, and observed data are often found to agree well with some SNLR model. For this reason, SNLR models are widely used by statisticians.

Sometimes a given regression experiment proves to be non-linear. In such cases, statisticians may seek an algebraic transformation of X which will yield a linear regression. Such a transformation leaves the assumptions of normality and homoscedasticity unaffected, since these assumptions apply only to the behavior of Y.

Bivariate experiments and bivariate models. If a probability experiment yields values for two random variables X and Y, it is called a bivariate experiment. The experiment of choosing an individual at random and measuring that person's height (in cm.) and weight (in kg.) is a bivariate experiment with X = height and Y = weight. If X and Y are continuous random

variables, a bivariate model is usually given by a joint probability density $h(x,y)$ as described in Chapter 16.

A bivariate experiment can be used as a regression experiment in the following way: To find an experimental value of Y for some given value x_0 of X, we simply conduct successive independent trials of the bivariate experiment, yielding $(x_1,y_1),(x_2,y_2),\ldots$ until we find a value x_i which is equal to x_0 or very close to x_0. We then use y_i as the value of Y observed for $X = x_0$.

Similarly, a bivariate model can be transformed into a regression model as follows. If x_0 is fixed, then as y varies, $h(x_0,y)$ measures the relative likelihood of observing y (for the given x_0). Hence

$$g(y|x_0) = \frac{h(x_0,y)}{\int_{-\infty}^{\infty} h(x_0,y)dy}$$

gives a probability density for Y. Then $g(y|x)$ is the desired regression model. When $g(y|x)$ is obtained from $h(x,y)$ in this way, it is sometimes called the conditional density obtained from $h(x,y)$ for Y given x. Note that

$$f(x) = \int_{-\infty}^{\infty} h(x,y)dy$$

will be the probability density for the random variable X when we consider X by itself as a single random variable. When $f(x)$ is obtained in this way, it is sometimes called the marginal density of X for the given bivariate model $h(x,y)$. Since

$$g(y|x) = \frac{h(x,y)}{f(x)} \quad , \quad \text{we also have}$$

$$h(x,y) = f(x)\, g(y|x) \quad .$$

This equation shows us how to go back from a regression model $g(y|x)$ to a bivariate model $h(x,y)$, provided that we know the density $f(x)$ for X.

If we begin with a bivariate model $h(x,y)$, if we then form $g(y|x)$ as above, and if we then find

$$E(Y|x) = \int_{-\infty}^{\infty} y\, g(y|x)\, dy, \quad \text{we call} \quad E(Y|x)$$

the <u>conditional expectation of</u> Y <u>given</u> x for the bivariate model $h(x,y)$.

We can also go from a bivariate model $h(x,y)$ for the random variables X and Y to a regression model in which X is the <u>dependent</u> variable and Y is the <u>independent</u> variable. For this, we use

$$f(x|y) = \frac{h(x,y)}{g(y)} \quad , \quad \text{where} \quad g(y) = \int_{-\infty}^{\infty} h(x,y)\, dx \, .$$

We then have the conditional expectation

$$E(X|y) = \int_{-\infty}^{\infty} x\, f(x|y)\, dx \quad .$$

To summarize: if we begin with a bivariate model $h(x,y)$, we can find regression models $g(y|x)$ and $f(x|y)$, as well as marginal densities $f(x)$ and $g(y)$, such that

$$h(x,y) = f(x)g(y|x) = g(y)f(x|y) .$$

Furthermore, the conditional expectations $E(Y|x)$ and $E(X|y)$ provide the regression equations $y = E(Y|x)$ and $x = E(X|y)$ for these regression models.

The transition from bivariate model to regression model is important because, as we shall see, we frequently begin with a bivariate experiement, then proceed to view it as a regression experiment, and then seek a good regression model for it. We discuss this further below.

Example. If we begin with the SNLR model for Y on X:

$$g(y|x) = N(y;x,1)$$

and then assume, in addition, that the parameter X is a random variable which follows the normal distribution

$$f(x) = N(x;0,1),$$

we obtain the bivariate model

$$h(x,y) = f(x)g(y|x) = \frac{1}{2\pi}e^{-\frac{1}{2}(2x^2-2xy+y^2)} .$$

We can now go from this bivariate model to a marginal density for Y. We obtain

$$g(y) = \int_{-\infty}^{\infty} h(x,y)dx = N(y;\ 0,\ \sqrt{2}),$$

as the reader may verify. Finally, as a regression model for X and Y, we obtain

$$f(x|y) = \frac{h(x,y)}{g(y)} = N(x;\tfrac{1}{2}y,\tfrac{\sqrt{2}}{2}) .$$

We shall see, later in this chapter, that it is always true that if X is normal and Y on X has an SNLR model, then Y is normal and X on Y has an SNLR model. Note also in this example that the regression curve for Y on X (given by the equation $y = x$) is distinct from the regression curve for X on Y (given by the equation $y = 2x$). This is true in general for regressions obtained from a continuous bivariate density.

Regression analysis. If we have data from a regression experiment, then a mathematical procedure for finding a regression model that agrees well with these observed data is called a regression analysis. If Y is dependent and X is independent, we speak of this as a regression analysis for Y on X. We shall now study certain forms of regression analysis for the case of SNLR models. The purpose in carrying out a regression analysis is to learn as much as we can about the way in which the dependent variable Y in a regression experiment depends upon the independent variable X. In each case, our hope is that this knowledge will give us an ability to make future predictions of values of Y from values of X and that it will lead us to a better understanding of the mechanism and natural laws which underlie the dependence of Y on X.

Independent trials. When we carry out a regression experiment, we choose a particular succession of values $x_1, \ldots, x_n$ for the independent variable and observe corresponding values $Y_1, \ldots, Y_n$ for the dependent variable. In using a regression model $g(y|x)$ as a mathematical picture of such an experimental procedure, we make two further assumptions about the relationship of the model to the experiment. These assumptions are as follows. (a) Given a specific value for x_i, the probabilistic behavior of Y_i is described by the density function $g(y|x_i)$ and is not affected by the values chosen

for $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$. (b) For each choice of $x_1, \ldots, x_n$, the variables $Y_1, \ldots, Y_n$ are independent (in the sense of independent random variables as defined in Chapters 16 and 17). When we use a regression model to describe a regression experiment, we refer to assumptions (a) and (b) together as the assumption of independence of trials.

In carrying out a regression analysis of a regression experiment, we shall always seek a regression model for which independence of trials holds. If we go from a bivariate experiment to a regression experiment, we will evidently have independence of trials, if trials of the bivariate experiment are independent. It often occurs in practice, however, that the successive trials of the bivariate experiment are not independent. In particular, the successive values of the independent variable X obtained in the bivariate experiment may represent some non-independent stochastic process. Under these circumstances, it is still possible for the derived regression experiment to have a regression model for which independence of trials holds, but we cannot be sure that this is so without further careful study. This difficulty is evident when we use bivariate historical data in a regression analysis. For example in economics, if we wish to do a regression analysis for Y = average per capita income on X = gross national product, the true relation between Y and X may be one in which the current year's value of Y is influenced not only by the current year's value of X, but by the values of X for recent years as well, and it may also be the case that for a given sequence of such X values, the corresponding Y values are not independent as random variables. Both parts of the assumption of independence of trials may thus fail. In spite of this failure, regression analysis is often carried out on such historical data. Evidently, the result of such an analysis will have limited predictive value and must be used with caution.

Estimating a regression model. The most common form of regression analysis is to go from observed data to the corresponding maximum-likelihood estimate for an SNLR model. Assume that we are given observed data in the form $(x_1,y_1),\ldots,(x_n,y_n)$, where (x_i,y_i) are the values of X and Y obtained on the

i^{th} trial of our given experiment. (In case the original experiment is bivariate, it is customary to treat all the observed values of X as if they were chosen parameter values for the regression experiment and hence to use all the data.) We take the universe U of possible regression models to be the set of all SNLR models. We formulate our maximum-likelihood problem as follows. Assume that the successive values $x_1,\ldots,x_n$ of the independent variable X are fixed. What model in U gives the maximum likelihood (with those specified $x_1,\ldots,x_n$) for the values $y_1,\ldots,y_n$ which were actually observed for Y.

The SNLR model which we seek must have the form $N(y;a+bx,c)$ for some values a,b,c. Evidently, by our independent trials assumption, the likelihood of our entire observation can be expressed as

$$L(y_1,\ldots,y_n;a,b,c) = N(y_1;a+bx_1,c)N(y_2;a+bx_2,c)\ldots N(y_n;a+bx_n,c)$$

$$= \frac{1}{(2\pi)^{n/2}c^n}\, e^{-\frac{1}{2}\sum_i \frac{(y_i - a - bx_i)^2}{c^2}}$$

Maximizing L is the same as maximizing $L' = \log L$, where $L' = -\frac{n}{2}\log 2\pi - n\log c - \sum_i \frac{1}{2c^2}(y_i - a - bx_i)^2$. To maximize L', we take partial derivatives with respect to a, b, and c and set these partial deriviatives = 0.

This gives:

$$\frac{\partial L'}{\partial a} = \frac{1}{c^2}\sum_i (y_i - a - bx_i) = 0;$$

$$\frac{\partial L'}{\partial b} = \frac{1}{c^2}\sum_i (y_i - a - bx_i)x_i = 0;$$

$$\frac{\partial L'}{\partial c} = -\frac{n}{c} + \frac{1}{c^3}\sum_i (y_i - a - bx_i)^2 = 0.$$

The first two equations become

$$y_i - na - b\sum x_i = 0$$

and

$$x_i y_i - a\sum x_i - b\sum x_i^2 = 0 .$$

If we introduce the notations $\bar{x} = \frac{1}{n}\sum x_i$, $\bar{y} = \frac{1}{n}\sum y_i$, $\overline{x^2} = \frac{1}{n}\sum x_i^2$, and $\overline{xy} = \frac{1}{n}\sum x_i y_i$, these two equations become

$$a + \bar{x}b = \bar{y}$$

and

$$\bar{x}a + \overline{x^2}b = \overline{xy} .$$

Solving for a and b, we have

$$a = \frac{\begin{vmatrix} \bar{y} & \bar{x} \\ \overline{xy} & \overline{x^2} \end{vmatrix}}{\begin{vmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{vmatrix}} = \frac{\overline{x^2}(\bar{y}) - \bar{x}(\overline{xy})}{\overline{x^2} - \bar{x}^2}$$

and

$$b = \frac{\begin{vmatrix} 1 & \bar{y} \\ \bar{x} & \overline{xy} \end{vmatrix}}{\begin{vmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{vmatrix}} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} ,$$

where the quantities on the right can be directly calculated from the data. The third equation (for $\frac{\partial L'}{\partial c} = 0$) now gives

$$c^2 = \frac{1}{n} \sum_i (y_i - a - bx_i)^2$$

in terms of a and b.

Note from the equation $a + \bar{x}b = \bar{y}$ that the estimated regression line $y = a + bx$ must go through the point $(\bar{x},\bar{y})$. Given this fact, the second equation $\bar{x}a + \overline{x^2}b = \overline{xy}$ then determines the slope b and intercept a of the regression line. The final equation gives c^2 as the average squared deviation of the data from this regression line.

Example. Assume that we have data from three trials: (1,0), (2,3), (3,3). (Usually, in a regression analysis, we will have data from a considerably larger number of trials.) We then calculate:

$$\bar{x} = \tfrac{1}{3}(1 + 2 + 3) = 2$$

$$\bar{y} = \tfrac{1}{3}(0 + 3 + 3) = 2$$

$$\overline{xy} = \tfrac{1}{3}(0 + 6 + 9) = 5$$

$$\overline{x^2} = \tfrac{1}{3}(1 + 4 + 9) = \tfrac{14}{3}.$$

Applying the formulas for a and b, we get

$$a = \frac{\frac{14}{3}(2) - 2(5)}{\frac{14}{3} - 4} = \frac{-2/3}{2/3} = -1$$

$$b = \frac{5 - 2(2)}{\frac{14}{3} - 4} = \frac{1}{2/3} = 3/2 .$$

Thus the regression for Y on X given by the maximum-likelihood SNLR model is

$$y = -1 + \frac{3}{2}x.$$

(This regression is also known as the least squares regression of Y on X because it is found by maximizing L with respect to a and b, which is the same as minimizing the sum of squares $\sum_i (y_i - a - bx_i)^2$.)

The maximum likelihood estimate for c is now found from

$$c^2 = \frac{1}{3}((0 + 1 - 3/2)^2 + (3 + 1 - 3)^2 + (3 + 1 - 9/2)^2)$$

$$= \frac{1}{3}(\frac{1}{4} + 1 + \frac{1}{4}) = 1/2.$$

We hence have $c = \frac{\sqrt{2}}{2}$.

What if we are given the same data, but wish to find a maximum-likelihood SNLR model with X as dependent variable and Y as independent variable? (This assumes that our data come originally from a bivariate experiment.) Then our regression will have the form $x = a' + b'y$ and we can find a' and b' by interchanging x and y in the formulas for a and b. This gives us

$$a' = \frac{\overline{y^2}(\bar{x}) - \bar{y}(\overline{xy})}{\overline{y^2} - \bar{y}^2}$$

and

$$b' = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{y^2} - \bar{y}^2}, \text{ where } \overline{y^2} = \frac{1}{n}\sum y_i^2.$$

From our data, $\overline{y^2} = \frac{1}{3}(0 + 9 + 9) = 6$. Hence we get

$$a' = \frac{6(a) - 2(5)}{6 - 4} = \frac{2}{2} = 1$$

and

$$b' = \frac{5 - 2(2)}{6 - 4} = \frac{1}{2},$$

and the maximum likelihood regression for X on Y is

$$x = 1 + \frac{1}{2}y.$$

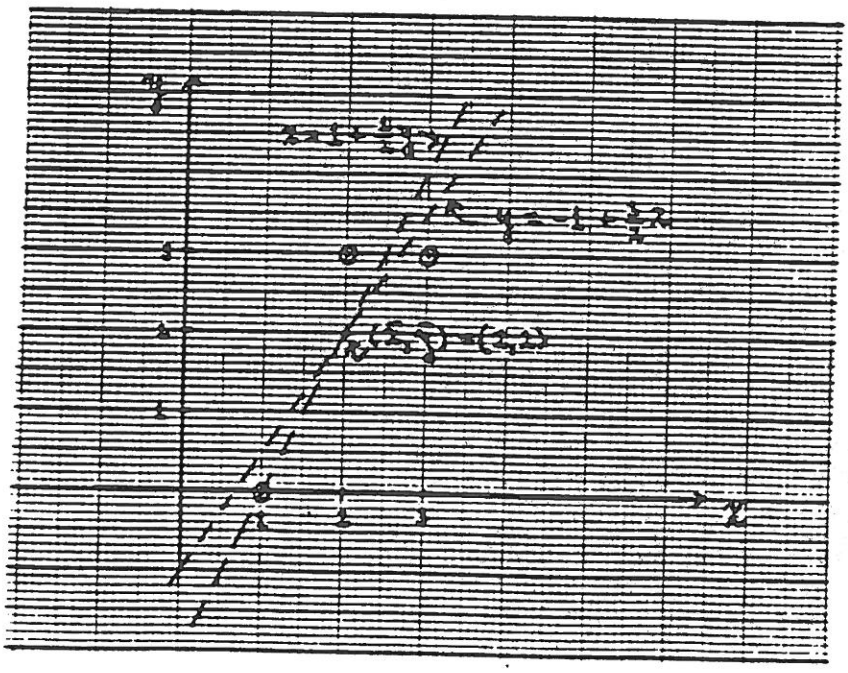
Similarly for the variance of the SNLR model for X on Y we have

$$c'^2 = \frac{1}{3}\sum (x_i - a' - b'y_i)^2$$

$$= \frac{1}{3}((1 - 1 - 0)^2 + (2 - 1 - \frac{3}{2})^2 + (3 - 1 - \frac{3}{2})^2)$$

$$= \frac{1}{3}(0 + \frac{1}{4} + \frac{1}{4}) = \frac{1}{6}, \quad \text{and} \quad c' = \frac{\sqrt{6}}{6}.$$

The two regression lines are shown, together with the data, in the following figure.



Note how each line corresponds, rather naturally, to the interpolated line that we would intuitively draw, from the data, for the desired dependence. (As we noted in a previous example, it is generally true that the regression line for Y on X is different from the regression line for X on Y. This is not surprising, since the regression of Y on X is concerned with a model for the behavior of Y as X varies, while the regression of X on Y is concerned with a model for the behavior of X as Y varies.

Causal relationships. Usually we carry out a regression analysis in order to find out more about how and why a dependent variable Y depends upon an independent variable X. In the case of a regression experiment in which the parameter X is under our direct control as successive trials of the experiment are carried out, we are likely to have good scientific grounds for believing that changes in the distribution of Y are directly caused by alterations in the chosen value of X. In the case of a regression experiment obtained from a bivariate experiment, however, such a causal relationship may be less obvious. Indeed, if we are not careful, a regression analysis of bivariate data may lead us unwittingly to conclude that there is a direct causal relationship when, in fact, none exists. The following example shows one regression for which a causal relationship exists and another regression for which a causal relationship does not exist.

Consider a bivariate experiment in which father-son pairs are randomly selected from a human population. X is defined

to be the adult height (in inches) of the father and Y is defined to be the adult height (in inches) of the son. We then consider this bivariate experiment as a regression experiment with X as independent and Y as dependent variable. An analysis of data from such an experiment provided one of the first published examples of a regression analysis (by Galton in 1903). The analysis was based on British data and used more than 1000 observed pairs. The estimated regression

$$y = 33.73 + .516x$$

was obtained. This result expresses the direct influence, in the observed population, of father's height on son's height. Evidently, the causal mechanism is largely genetic.

The same bivariate experiment, however, can also be considered as a regression experiment with Y as independent and X as dependent variable. Using the same data, the equation

$$x = 32.53 + .512y$$

is obtained for the regression of X on Y. Obviously, this record result cannot be viewed as describing a direct causal relationship, since the height of a son is determined at a later time than the height of his father. For this reason, the regression of X on Y is perhaps of less interest in this experiment than the regression of Y on X. (On the other hand, if we know the son's height but not the father's, we <u>can</u> use this regression equation of X on Y to help us

to make a good guess as to the father's height.)

Regression analyses based on data from bivariate experiments occasionally become subjects of scientific controversy because of doubts as to the nature (or even the existence) of a direct causal relationship between the independent and dependent variables. Such controversies can be both subtle and complex. For example, it may be argued that a correct causal interpretation of the influence of X on Y must involve some third unexamined variable that is directly related to both X and Y. In fact, a third relevant variable does occur in the case of father-son heights and must be considered if we are fully to understand the way in which X influences Y. This variable is the mother's height. Further analysis shows that the causal influence of the father's height not only includes the father's direct genetic influence but also includes a tendency of taller men to marry taller women and shorter men to marry shorter women, so that, in the regression of Y on X, the father's height acts indirectly through the mother's genetic influence as well as directly through the father's own genetic influence.

We note two other special features of the father-son example. First, observe that the regression equation for Y on X ($y = 33.73 + .516x$) is of the form $y = a + bx$ with $b < 1$. This means that, in the observed population, taller-than-average fathers tend to have sons who are shorter than they are, and shorter fathers tend, on the average, to have taller sons. In the 1903 analysis, this tendency was spoken of as a "regression" of the son's height <u>from</u> the father's height <u>towards</u> the population average. We call such a tendency

a regression effect.

Second, observe that the regression equation for X on Y ($x = 33.53 + .512y$) also shows a regression effect. Taller-than-average sons tend to have had fathers who are shorter than they are, and shorter sons tend to have had taller fathers. This regression effect from sons to fathers does not contradict the previously described regression effect from fathers to sons. The previous effect relates to variations in Y for fixed x, while the present effect relates to variations in X for fixed y. We shall see later in this chapter that these two regression effects do not indicate anything unusual about the father-son experiment and data, but are, instead, a typical mathematical feature of experiments and data of this general kind. In fact, the word "regression" is now used broadly to describe the forms of model, experiment, and analysis that we have already called regression models, regression experiments, and regression analyses.

Hypothesis tests and confidence regions for SNLR models. In previous chapters, we have defined a model to be a probability space. In the present chapter, we are using "model," in the sense of regression model, to mean a parametrized family of probability spaces. Once we have chosen a particular sequence of values for the independent variable in a regression experiment, the concepts of metric, hypothesis test, and confidence region can be used with regression models in much the same way as before. We do this now for the case of SNLR models.

Assume that $N(y;\alpha+\beta x,\sigma)$ is a given SNLR model with fixed α, β, and σ, and assume that we have a regression experiment for which this is the correct model. If we fix values of $x_1,\ldots,x_n$ of X and observe corresponding values $y_1,\ldots,y_n$ of Y, we can proceed to calculate a, b, and c as described above. The quantities $\bar{y}$, a, b, and c may be viewed as new random variables whose distributions (for the given fixed $x_1,\ldots,x_n$) depend upon α, β, and σ. What can we say about the nature and form of these distributions?

We first observe that $\bar{y} = \frac{1}{n}\sum_i y_i$ is a sum of independent normal variables and hence must have a normal distribution with

$$E_{\bar{y}} = \frac{1}{n}\sum_i (\alpha+\beta x_i) = \alpha + \beta\bar{x}$$

and

$$V_{\bar{y}} = \sigma^2/n.$$

Similarly, $$b = \frac{\overline{xy} - \bar{x}(\bar{y})}{\overline{x^2} - \bar{x}^2} = \frac{1}{\overline{x^2} - \bar{x}^2}\left(\sum_i \frac{x_i y_i}{n} - \bar{x}\sum_i \frac{y_i}{n}\right)$$

$$= \frac{1}{n(\overline{x^2} - \bar{x}^2)}\sum_i (x_i - \bar{x})y_i$$

is a sum of independent normal variables and hence must be normal. (Note that the x_i are all fixed, so that the y_i are the only random variables appearing in this expression.) It is easy to show, by the algebraic laws of expectation and variance, that

$$E_b = \beta$$

and

$$V_b = \frac{\sigma^2}{n(\overline{x^2} - \bar{x}^2)}.$$

Similarly, one may show that $a = \dfrac{\overline{x^2}(\bar{y}) - \bar{x}(\overline{xy})}{\overline{x^2} - \bar{x}^2}$ is normal with

$$E_a = \alpha$$

and

$$V_a = \frac{\sigma^2 \overline{x^2}}{\overline{x^2} - \bar{x}^2} .$$

Finally one may show that $\dfrac{n(c^2)}{\sigma^2}$ has a chi-square distribution with $n-2$ degrees of freedom. (See Chapter 18.)

For a given SNLR model with $\beta = \beta_0$, and for a given fixed choice of $x_1,\ldots,x_n$, we introduce the metric

$$s(\omega) = \frac{(b - \beta_0)}{c} ,$$

where ω is an observation $(y_1,\ldots,y_n)$. Using the methods of Chapter 18, we can show that this metric is well defined (for the fixed choice of $x_1,\ldots,x_n$) on the composite model consisting of all SNLR models with $\beta = \beta_0$ (where α and σ may vary). We can also show that

$$s(\omega)\sqrt{(n-2)(\overline{x^2} - \bar{x}^2)} = \frac{b - \beta_0}{c}\sqrt{(n-2)(\overline{x^2} - \bar{x}^2)}$$

follows a t distribution with $n-2$ degrees of freedom. This enables us to calculate DLS values for our metric and hence to carry out hypothesis tests for a given β_0 or to construct confidence intervals for β. In particular, if we take $\beta_0 = 0$, we have a hypothesis test for the null hypothesis that Y does not depend on X.

Other metrics for testing or approximating values of α and values of σ can be defined and used in similar ways. It is also possible to define a metric which is well-defined on the set of all SNLR models (considered as a composite model in the universe of all normal, homoscedastic, regression models) and to use this metric to carry out a hypothesis test for the composite null hypothesis that the regression is linear. Further details on these tests and constructions may be found in more advanced texts on regression methods.

Example. We use the data in the previous example to find a 95% confidence interval for β. We had $b = 3/2$, $c^2 = 1/2$, $n = 3$, $\bar{x} = 2$, and $\overline{x^2} = 14/3$. From our t-table, we find that the 95% points for a t-distribution with 1 degree of freedom occur at $t = \pm 12.7$. Hence we have

$$\frac{\frac{3}{2} - \beta}{\sqrt{\frac{1}{2}}} \sqrt{\frac{2}{3}} = \pm 12.7$$

or

$$= \frac{3}{2} \pm 12.7\sqrt{\frac{4}{3}} = \frac{3}{2} \pm 14.7.$$

Thus our confidence interval for the slope β of linear regression is (-13.2,16.2). (The interval is so large because the number of observed data points is so small.)

Example. In the father-son experiment described above, 1078 pairs were observed, and the values $\overline{x^2} - \bar{x}^2 = 7.29$ and $c = 2.32$ were obtained. We find a 95% confidence interval for β as follows. Since the t distribution with 1076

degrees of freedom is extremely close to the standard normal distribution, we have

$$\frac{.516 - \beta}{2.32}\sqrt{(1076)7.29} = \pm 1.96.$$

Hence $$\beta = .516 \pm \frac{2.32(1.96)}{\sqrt{(1076)7.29}} = .516 \pm .051.$$

Multiple regression. Regression experiments and regression models can be defined in an exactly similar way, when we wish to relate the dependent variable Y to several independent parameter variables X,Z,W,... at the same time. Such experiments and models are said to be multiple regression experiments and models. (The case of a single independent variable is usually called simple regression.) The derivations and calculations for multiple regression are similar to the case of simple regression described above. (See the Exercises.) As before, normal linear homoscedastic regression models are widely used. As before, the most common form of multiple regression analysis is a maximum likelihood "least squares" calculation.

Multiple regression analysis can be especially helpful in exploring and identifying causal relationships, since they give us a way to measure and compare the relative simultaneous effects of different independent variables upon the same dependent variable. For example, in the father-son height experiment described earlier, information was also gathered on mothers' heights, and these data were used to carry out a multiple regression analysis. This resulted in the regression

$$y = 14.08 + .409x + .430z,$$

where y = son's height, x = father's height, and z = mother's height. It is instructive to compare such an observed regression with regressions that one might expect to obtain on the basis of various given scientific assumptions. Regression comparisons of this kind can provide important scientific insights and suggest valuable new directions for research. This is especially true in cases (like that of son-father-mother heights) where the observed data are found to follow some linear regression model extraordinaly well. We consider these matters further at the end of the present chapter.